C-005

ストレージシステムのリモートコピー機能におけるスケールアウト化手法の提案と評価
Study on Scale-out Methods for Remote Copy Function in Storage System

原 彬大 †　出口 彰 †　山本 大輔 †
Akihiro Hara　Akira Deguchi　Daisuke Yamamoto

1. はじめに

災害等に伴い企業の情報システムが停止すると巨額の損失を招く。このため、遠隔地に災害対策用サイトを設け、災害時には災害対策用サイトに切り替えて業務を継続するディザスタリカバリが一般的となっている。このディザスタリカバリを支援する機能として、ストレージシステム(以降、ストレージと記載)においては、ホストから書き込まれたデータを非同期で災害対策用サイトにコピーするリモートコピー機能を提供してきた。

近年では、IT システムのアジリティ向上が求められており、小規模なシステムを構築し必要に応じて規模拡大していくスモールスタートな運用が増えている。規模に応じてストレージが追加され複数のストレージにデータが分散配置される場合があることから、複数ストレージに跨ったリモートコピーが求められる。一方で、災害発生時に災害対策用サイトで業務を再開するためには、正サイトでホストがストレージに書き込んだ関連データ全ての書き込み順序を維持しつつリモートコピーを行う必要がある[1]。本研究では、複数ストレージに跨ってデータの書き込み順序を維持する従来のコピー方式を拡張し、規模拡大時に発生するホスト I/O のスループットの低下を抑制する方式を提案し、提案方式適用前後のスループットをモデル化し評価した。

2. 複数ストレージに跨った非同期リモートコピーの従来方式と問題点

2.1 従来方式

本節では、ストレージ 1 台で性能が不足する場合向けに、複数ストレージに跨ってデータの書き込み順序を維持しつつ非同期リモートコピーを可能にする従来方式[2]について図 1 を参照しつつ説明する。

まず、ストレージ間の非同期リモートコピーにおけるデータの流れについて図 1 中の①～④を用いて示す。

① ホストが LU(Logical Unit)に対してデータ書き込み指示を出す。
② ホストからの書き込みと同期して、データと書き込み順序を示す情報を持つジャーナル(以降 JNL と記載)が作成され、JNL VOL に格納される。
③ 作成された JNL がコピー先ストレージの JNL VOL にネットワークを経由し転送される。
④ JNLVOL から JNL を読み出し、JNL に保持されている書き込み順序を示す情報を参照して、ホストから書き込まれた順序を維持しつつコピー先ボリュームに書き込みデータを反映する。

従来方式では、②に関連する複数ストレージに跨ってデータの同期点を作成するデータ同期点作成機能と、④に関連する同期点を基にコピー先ボリュームへのデータ反映を制御するデータ反映調整機能の 2 つの機能によって複数ストレージに跨ってデータの書き込み順序を維持したコピーを実現している。各機能について順に説明する。

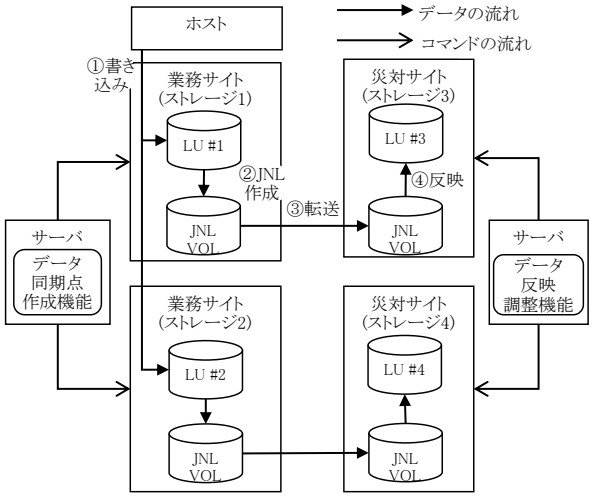

図 1 複数ストレージに跨ったリモートコピーの概要図

データ同期点作成機能はある一ヶ所で動作し、書き込み処理を停止・再開する指示と、書き込み順序を示す世代番号を更新する指示を、関連する複数のコピー元ストレージに発行する。具体的な動作手順としては、ホストからの書き込み処理を停止するコマンドを関連するコピー元ストレージ全てに順に実行し、全ての完了を受け取る。次に書き込み処理の再開指示と共に JNL に付与する書き込み順序を示す世代番号を全ストレージに順に配布する。この世代番号はホストからの書き込みを受領し JNL を作成する時に書き込み順序を示す情報として付与される。この動作によって、各同期点毎に前回の同期点から書き込まれたデータを複数ストレージ間で同一の世代として管理できる。

データ反映調整機能はある一ヶ所で動作し、到着済世代番号を確認するとともに、ある同期点までのデータ反映を関連するコピー先ストレージに対して指示する。具体的な動作手順としては、関連するコピー先ストレージ全てから最新の JNL の世代番号を取得する。次に、ある同期点に含まれるデータが全て揃っていることが保証されるデータとして、全てのストレージに共通して到着している世代番号から一つ前の世代番号を持つデータまでを一度にまとめてコピー先ボリュームに反映する指示を出す。この動作により、複数ストレージに跨って書き込み順序が揃っているデータ断面毎にコピーが可能になる。

2.2 従来方式の問題点

データ同期点作成機能はホストからの書き込み停止を定期的に実施する。このときに、書き込みが停止する時間に応じてホスト I/O のスループットが低下してしまう。従来方式においては全ストレージに順番に指示を出す逐次実行

となっており、関連するストレージの数が増えるに従いホスト I/O のスループットが低下する。この特性は近年のスモールスタートの潮流に伴いストレージの数が増えていることを考慮すると問題になると考えられる。

3. 提案方式

ホスト I/O のスループット低下を防ぐためには、ホストの書き込みを停止するデータの同期点作成にかかる時間を短縮すればよい。データの同期点作成は、各ストレージが書き込みを停止し、全ストレージの完了を待ち合わせてから世代番号を付与し、書き込みを再開する順番で実行される。このとき、データ同期点作成機能が行う各ストレージへの指示についてはストレージ毎の逐次実行は必須でないため並列実行しても書き込み順序維持に影響はない。そこで、本稿においてはデータの同期点作成処理が関連するコピー元ストレージ全てに対して指示を並列で発行する方式を提案する。本方式によって、リモートコピーを構成するストレージの数によらずホスト I/O 停止時間が一定となりホスト I/O のスループット低下抑制が期待できる。

4. 評価

4.1 ホスト I/O のスループットのモデル化と評価方法

ホスト I/O のスループットが最大限低下する I/O パターンである書き込みのみのケースにおいて、リモートコピーを構成するストレージの数とホスト I/O のスループットの関係をモデル化し見積もる。ホスト I/O のスループット低下割合(R_{loss})は 1 回のデータ同期点作成の間隔(T_{freeze})のうち、ホスト I/O 停止時間が占める割合と一致する。データ同期点作成 1 回あたりのホスト I/O 停止時間は、データ同期点作成機能とコピー元ストレージの間のネットワークの転送時間($T_{network}$)とストレージがデータ同期点作成機能からの指示を受領してから実行完了するまでの処理時間($T_{process}$)の和となる。また、従来方式では指示がストレージ毎に逐次実行されるためこの和に対してストレージ数($N_{storage}$)を乗算し計算する。以上より従来方式におけるホスト I/O のスループット低下割合は計算式(1)で表される。

$$R_{loss} = \frac{(T_{network} + T_{process})N_{storage}}{T_{freeze}} \quad (1)$$

本稿で提案する並列発行方式におけるホスト I/O のスループット低下割合(R_{loss})はストレージ数($N_{storage}$)に依存しないため、計算式(2)で表される。

$$R_{loss} = \frac{(T_{network} + T_{process})}{T_{freeze}} \quad (2)$$

ホスト I/O のスループット維持率($R_{retention}$)は、各方式の低下率の余りの割合であり計算式(3)で表される。

$$R_{retention} = (1 - R_{loss}) \quad (3)$$

リモートコピー全体のホスト I/O のスループット(W)は各方式のストレージあたりのホスト I/O スループット($W_{storage}$)とスループット維持率($R_{retention}$)とストレージ数($N_{storage}$)の乗算で計算でき、計算式(4)により表される。

$$W = W_{storage}R_{retention}N_{storage} \quad (4)$$

†株式会社 日立製作所 研究開発グループ　Hitachi, Ltd., Research & Development Group

4.2 評価結果

一般的にストレージへの書き込み応答時間は 0.5 ms 程度であり、ストレージ内の処理時間を書き込み応答時間同等の 0.5 ms 程度と仮定する。ネットワーク転送時間はデータ同期点作成機能の動作位置とコピー元ストレージ間の距離に依存するため変数とする。表 1 に示すパラメータを従来の逐次実行方式と提案する並列発行方式に当てはめ、式(1)と式(2)と式(3)よりストレージ数毎にホスト I/O のスループット維持率を計算した結果を図 2 に示す。

表 1. 見積もりのパラメーター一覧

#	項目	時間 [ms]
1	データ同期点作成の実行間隔	1,000
2	ネットワークの転送時間	0, 5, 10
3	ストレージ内の処理時間	0.5

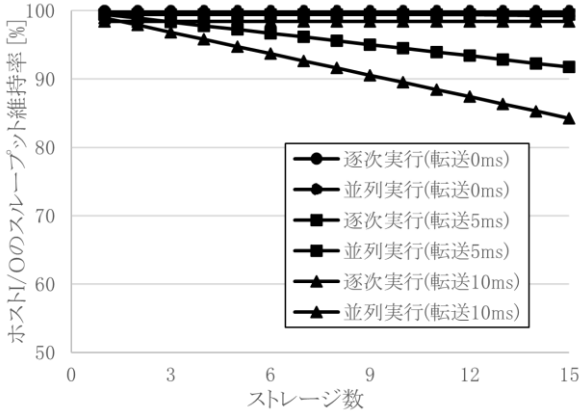

図 2 ストレージ数とホスト I/O スループット維持率の関係

図 2 より、従来方式ではネットワークの転送時間が小さくストレージ数が 15 台までの構成ではホスト I/O のスループット低下は軽微であった。しかし、転送時間が延びるとストレージ数に応じたスループット低下が見られることが分かった。一方で提案方式ではネットワークの転送時間によらずホスト I/O のスループット低下が軽微だと分かった。

4.3 考察

従来の逐次実行方式は実装がシンプルなメリットがある。ただし、データ同期点作成機能の動作位置とコピー元のストレージ間のネットワークが離れているケースにおいてはストレージ数に応じてホスト I/O スループットが低下する問題があることが分かった。このため、例えばスモールスタートな運用を行うためにストレージ数が多く、またストレージ間の物理的な距離が離れている環境においては本稿で提案する並列発行方式が適していると考えられる。

5. まとめ

本稿では、複数ストレージ間の非同期リモートコピー機能において規模拡大時に発生するホスト I/O スループット低下を抑制する方式を提案した。これにより、ストレージ数やデータセンタの物理位置等によらずディザスタリカバリシステムのホスト I/O スループットを維持可能となった。

参考文献

[1] 出口, 二瀬, et al, “大規模ディザスタリカバリシステム向け非同期リモートコピーの研究”, FIT 2006, (2006).
[2] 日立製作所, “Raid Manager ユーザガイド”, (2021).